

BIROn - Birkbeck Institutional Research Online

Gong, C. and Tao, D. and Maybank, Stephen J. and Liu, W. and Kang, G. and Yang, J. (2016) Multi-modal curriculum learning for semi-supervised image classification. IEEE Transactions on Image Processing 25 (7), pp. 3249-3260. ISSN 1057-7149.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/15093/>

Usage Guidelines:

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>
contact lib-eprints@bbk.ac.uk.

or alternatively

Multi-modal Curriculum Learning for Semi-supervised Image Classification

Chen Gong, Dacheng Tao, *Fellow, IEEE*, Stephen J. Maybank, *Fellow, IEEE*, Wei Liu, *Member, IEEE*, Guoliang Kang, and Jie Yang

Abstract—Semi-supervised image classification aims to classify a large quantity of unlabeled images by harnessing typically scarce labeled images. Existing semi-supervised methods often suffer from inadequate classification accuracy when encountering difficult yet critical images such as outliers, because they treat all unlabeled images equally and conduct classifications in an imperfectly ordered sequence. In this paper, we employ the curriculum learning methodology by investigating the difficulty of classifying every unlabeled image. The reliability and discriminability of these unlabeled images are particularly investigated for evaluating their difficulty. As a result, an optimized image sequence is generated during the iterative propagations, and the unlabeled images are logically classified from simple to difficult. Furthermore, since images are usually characterized by multiple visual feature descriptors, we associate each kind of features with a “teacher”, and design a Multi-Modal Curriculum Learning (MMCL) strategy to integrate the information from different feature modalities. In each propagation, each teacher analyzes the difficulties of the currently unlabeled images from its own modality viewpoint. A consensus is subsequently reached among all the teachers, determining the currently simplest images (*i.e.* a curriculum) which are to be reliably classified by the multi-modal “learner”. This well-organized propagation process leveraging multiple teachers and one learner enables our MMCL to outperform five state-of-the-art methods on eight popular image datasets.

Index Terms—Curriculum learning, Semi-supervised learning, Multi-modal, Image classification.

I. INTRODUCTION

CLASSIFYING natural images into meaningful categories has always been a dominant topic in computer vision research. With the emergence of large image collections and

vigorous development of the Internet, the available labeled images are usually inadequate for training a supervised classifier that has to deal with a dramatic growth in new images. Furthermore, labeling more images will incur high time and monetary costs. To address this problem, semi-supervised image classification has been developed to explicitly exploit the information revealed by both limited labeled images and sufficient unlabeled images [1], [2], [3], [4], [5], [6], [7].

Given labeled image set \mathcal{L} of size l and unlabeled image set \mathcal{U} of size u , conventional semi-supervised image classification is usually conducted on a weighted similarity graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ [1], [8], [9], where \mathcal{V} is the node set representing the $n = l + u$ images, and \mathcal{E} is the edge set encoding the pairwise similarities between these images. The target is to iteratively propagate the labels from \mathcal{L} to \mathcal{U} so that all the elements in \mathcal{U} can be precisely classified. However, existing methods [1], [8], [9] often yield unsatisfactory results, as they are very likely to make incorrect classifications on “outliers” or “bridge examples” (*e.g.* images that share similar properties with multiple classes). This is because existing methods treat all the unlabeled images equally without considering the difficulty or reliability of their classification. As a result, the images are classified in imperfect order, which leads to the error-prone classification of difficult but critical images such as the aforementioned “outliers” and “bridge examples”. Such errors have an adverse impact on the accurate prediction of the labels of the remaining unlabeled images.

Based on the above consideration, we assume that different images have different levels of difficulty, and utilize curriculum learning [10] to re-organize the learning sequence (*i.e.* the classification order for the unlabeled images), so that the unlabeled images are logically classified from simple to difficult. As a result, the unlabeled images can be reliably labeled because this well-organized learning sequence enables the previously attained simple knowledge to facilitate the subsequent classification of complex images. Taking account of the fact that an image can usually be characterized by different feature descriptors, we regard each type of features as one modality and develop “Multi-Modal Curriculum Learning” (MMCL) to guide the learning process. As a result, the consistency and complementarity of various features can be fully exploited. Our MMCL strategy is very similar to the human’s acquisition of knowledge during the various stages from childhood to adulthood, during which time an individual gains knowledge from many teachers of different subjects. These different subjects are naturally analogous to the different feature modalities in our algorithm.

This research is supported by NSFC of China (No: 61572315), 973 Plan of China (No. 2015CB856004), and Australian Research Council Projects DP-140102164, FT-130101457, and LE-140100061.

C. Gong is with the Institute of Image Processing and Pattern Recognition in Shanghai Jiao Tong University. He is also with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia. (e-mail: goodgongchen@sjtu.edu.cn).

D. Tao and G. Kang are with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au; kgl.pml@gmail.com).

S. Maybank is with the School of Computer Science and Information Systems, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK. (email: sjmaybank@dc.s.bbk.ac.uk).

W. Liu is with the IBM T. J. Watson Research Center, 1101 Route 134 Kitchawan Rd, Yorktown Heights, NY 10598, USA (e-mail: wliu@ee.columbia.edu).

J. Yang is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China, 200240 (e-mail: jieyang@sjtu.edu.cn).

Corresponding authors: Jie Yang (jieyang@sjtu.edu.cn) and Dacheng Tao (dacheng.tao@uts.edu.au).

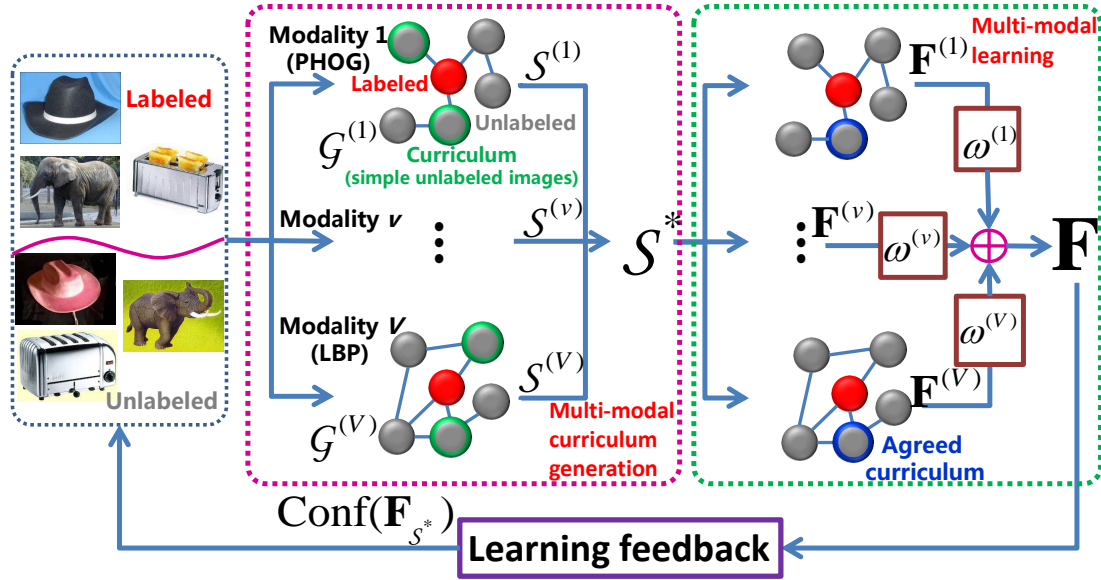


Fig. 1. The framework of our algorithm. All the labeled and unlabeled images are described by V different modalities (*i.e.* features). The labeled images and unlabeled images are represented by red and grey balls, respectively. Among the unlabeled images, the curriculum of each modality and the agreed curriculum S^* are further surrounded by green and blue rings accordingly. The multi-modal curriculum generation step and multi-modal learning step are marked with magenta and green dashed boxes.

In detail, we build V graphs $\mathcal{G}_1, \dots, \mathcal{G}_V$ corresponding to V modalities over the images in $\mathcal{L} \cup \mathcal{U}$ (see Fig. 1). The state-of-the-art label propagation approach [11] is applied to these V modalities to form a multi-modal semi-supervised learner, which can iteratively classify the unlabeled images. Besides, multiple “teachers” are incorporated to allocate the simplest unlabeled images to this stepwise semi-supervised learner. Here the simplest unlabeled images constitute a curriculum as they should be “learned” by the learner as required by the teachers. In each propagation, the difficulties of unlabeled images (grey balls) are evaluated by all V teachers according to their reliability and discriminability w.r.t. the labeled images (red balls), and the curriculum images decided by every individual teacher are denoted by the set $\mathcal{S}^{(v)}$ ($v = 1, \dots, V$) (gray balls with green rings). An optimal curriculum S^* (gray balls with blue rings) agreed by all the teachers is then established based on every individual teacher’s decision. After that, the learner classifies the images in S^* by respectively propagating the labels from \mathcal{L} to S^* in V different modalities, and the obtained results are recorded in the label matrix $\mathbf{F}^{(v)} \in \mathbb{R}^{n \times c}$ ($v = 1, \dots, V$, and c is the number of classes). This is also known as multi-modal learning. The integrated label matrix $\mathbf{F} \in \mathbb{R}^{n \times c}$ is the sum of $\mathbf{F}^{(1)}, \dots, \mathbf{F}^{(V)}$ weighted by $\omega^{(1)}, \dots, \omega^{(V)}$, respectively. Finally, the learning feedback is delivered to teachers to assist them to correctly determine the subsequent simplest curriculum. This process iterates until all the images in \mathcal{U} have been selected, and the label matrix thus produced is denoted by $\bar{\mathbf{F}}$. The (i, j) -th element in $\bar{\mathbf{F}}$ (or $\mathbf{F}^{(v)}$ and \mathbf{F} mentioned above) encodes the probability of the i -th image \mathbf{x}_i belonging to the j -th class \mathcal{C}_j .

Due to the wisdom of multiple teachers, the difficulties of unlabeled images are comprehensively evaluated, and these images are logically propagated from simple to difficult with

the updated curriculums. Our algorithm consequently achieves higher classification accuracy than other typical methods, as revealed by empirical validation. The advantage of multi-modal curriculum over single-modal curriculum is also demonstrated in the experiments.

II. RELATED WORK

In this section, we review some representative existing literatures of semi-supervised image classification, multi-modal learning and curriculum learning, as they are related to this work.

A. Semi-supervised Image Classification

Semi-supervised learning (SSL) [12] has been studied for a long history, which aims to classify a massive number of unlabeled examples given the existence of only a few labeled examples. Although the massive unlabeled examples do not have explicit labels, they convey the distribution information of the entire dataset, which can be exploited for accurate classification. Existing SSL algorithms can be roughly attributed to three categories: self-training [13], low density separation [14], [15], [16], and graph-based methods [17], [18], [19].

With respect to its application to image classification, Dai et al. [2] proposed to learn better image representation with the aid of the available image data. To be specific, they sample an ensemble of image prototypes from both labeled and unlabeled images, and then learn a discriminative feature representation of an unlabeled image by computing its projected values on the previously sampled prototypes. Shrivastava et al. [20] deployed semi-supervised bootstrapping to gradually classify the unlabeled images in a self-learning way. In this work, the semi-supervised learning is constrained by the common attributes shared across different classes as well as the attributes

which make one class different from another. Fergus et al. [3] developed a scalable graph-based algorithm that has linear complexity with regard to the number of images. The spectral property of the graph is properly utilized to handle the large-scale image data. Other representative works include [5], [21], [22], [23].

B. Multi-modal Learning

In practical applications, data is often obtained from multiple sources rather than a single source. One common way to process such multi-modal data is to concatenate the feature vectors associated with different sources into a long vector. However, this concatenation is highly intuitive and ignores the particular statistical property of an individual view, therefore it can hardly obtain satisfactory performance. Multi-modal learning (MML) is therefore proposed to explicitly fuse the complementary information from different modalities to achieve improved performance. MML is essentially multi-view learning, and the relevant algorithms can be classified into three groups [24]: co-training [25], [26], multiple kernel learning [27], and subspace learning [28]. Co-training approaches train alternately to maximize the mutual agreement on different modalities of the unlabeled data. Multiple kernel learning straightforwardly corresponds to multiple modalities and elegantly combines kernels of different modalities to achieve improved performance. Subspace learning assumes that there is a latent subspace shared by multiple modalities and the input modalities are generated from this low-dimensional latent subspace.

MML has been widely adopted in semi-supervised image classification such as multi-modal SSL [4], adaptive multi-modal SSL [1], and multi-view vector-valued manifold regularization [29]. Of these, [4] is based on a multiple kernel classifier fusing both image content and its descriptive keywords. The work of [1] integrates various heterogeneous visual features via graph fusion, and then deploys label propagation [18] to infer the class labels of unlabeled images. By utilizing the vector-valued functions, [29] proposes a multi-modal algorithm for multi-label image classification.

C. Curriculum Learning

Curriculum learning aims to improve the learning performance by designing suitable curriculums from simple to difficult for the stepwise learner. This learning approach was proposed by [10], which hypothesizes that curriculum learning is able to boost the convergence speed of the training process as well as find a better local minima than the existing solvers for non-convex problems. The self-paced learning proposed by Kumar et al. [30] can be regarded as an implementation of curriculum learning, which was extended by [31], [32] afterwards. Besides, the teaching-to-learn and learning-to-teach framework developed by Gong et al. [33], [34] also follows the idea of curriculum learning, and this paper is the extension of [33] to the multi-modal situation.

Up to now, curriculum learning has been applied to visual category discovery [35], object tracking [36], and multimedia

retrieval [37]. However, none of the existing curriculum learning algorithms can handle multi-modal data, or touch semi-supervised image classification.

III. OUR APPROACH

For each feature modality (indexed by $v = 1, \dots, V$), we construct a similarity graph $\mathcal{G}^{(v)}$ by the recently proposed adaptive edge weighting [38], and the associated adjacency matrix is $\mathbf{W}^{(v)}$ with the (i, j) -th element $\mathbf{W}_{ij}^{(v)}$ representing the similarity between images \mathbf{x}_i and \mathbf{x}_j in terms of modality v . The graph Laplacian is $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{W}^{(v)}$ where $\mathbf{D}^{(v)}$ is the degree matrix with diagonal elements computed by $\mathbf{D}_{ii}^{(v)} = \sum_{j=1}^n \mathbf{W}_{ij}^{(v)}$.

A. Single-modal Curriculum Generation

We start by elaborating the curriculum generation on single feature modality, so the superscript (v) in previous notations is temporarily dropped in this section. The purpose of curriculum generation is to pick up the simplest curriculum $\mathcal{S} \subset \mathcal{U}$ in each propagation. To this end, the *reliability* and *discriminability* of every unlabeled image are investigated by the “teacher” to make a selection.

Specifically, we assign a random variable y_i to each image \mathbf{x}_i , and view the propagations on \mathcal{G} as a Gaussian process [39]. Therefore, this Gaussian process is modeled as a multivariate Gaussian distribution over the random variables $\mathbf{y} = (y_1, \dots, y_n)^\top$, which has a concise form $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with its covariance matrix being $\mathbf{\Sigma} = (\mathbf{L} + \mathbf{I}/\kappa^2)^{-1}$ (\mathbf{I} is the identity matrix). The parameter κ^2 controls the “sharpness” of the distribution and is fixed at 100 throughout this paper.

A curriculum \mathcal{S} is reliable w.r.t. the labeled set \mathcal{L} if the conditional entropy $H(\mathbf{y}_{\mathcal{S}}|\mathbf{y}_{\mathcal{L}})$ is small, where $\mathbf{y}_{\mathcal{S}}$ and $\mathbf{y}_{\mathcal{L}}$ denote the subvectors of \mathbf{y} corresponding to \mathcal{S} and \mathcal{L} , respectively. This is because small $H(\mathbf{y}_{\mathcal{S}}|\mathbf{y}_{\mathcal{L}})$ suggests that the curriculum set \mathcal{S} comes as no “surprise” to the labeled set \mathcal{L} . Besides, a curriculum is discriminable if the included images are significantly inclined to certain classes.

Reliability. By using the property of multivariate Gaussian [40], the most reliable curriculum can be found by optimizing:

$$\begin{aligned} & \min_{\mathcal{S} \subset \mathcal{U}} H(\mathbf{y}_{\mathcal{S}}|\mathbf{y}_{\mathcal{L}}) \\ & \Leftrightarrow \min_{\mathcal{S} \subset \mathcal{U}} H(\mathbf{y}_{\mathcal{S} \cup \mathcal{L}}) - H(\mathbf{y}_{\mathcal{L}}) \\ & \Leftrightarrow \min_{\mathcal{S} \subset \mathcal{U}} \left(\frac{s+l}{2} (1 + \ln 2\pi) + \frac{1}{2} \ln |\mathbf{\Sigma}_{\mathcal{S} \cup \mathcal{L}, \mathcal{S} \cup \mathcal{L}}| \right) \\ & \quad - \left(\frac{l}{2} (1 + \ln 2\pi) + \frac{1}{2} \ln |\mathbf{\Sigma}_{\mathcal{L}, \mathcal{L}}| \right) \\ & \Leftrightarrow \min_{\mathcal{S} \subset \mathcal{U}} \frac{s}{2} (1 + \ln 2\pi) + \frac{1}{2} \ln \frac{|\mathbf{\Sigma}_{\mathcal{S} \cup \mathcal{L}, \mathcal{S} \cup \mathcal{L}}|}{|\mathbf{\Sigma}_{\mathcal{L}, \mathcal{L}}|}, \end{aligned} \quad (1)$$

where $\mathbf{\Sigma}_{\mathcal{L}, \mathcal{L}}$ and $\mathbf{\Sigma}_{\mathcal{S} \cup \mathcal{L}, \mathcal{S} \cup \mathcal{L}}$ are submatrices of $\mathbf{\Sigma}$ associated with the corresponding subscripts. By further partitioning $\mathbf{\Sigma}_{\mathcal{S} \cup \mathcal{L}, \mathcal{S} \cup \mathcal{L}} = \begin{pmatrix} \mathbf{\Sigma}_{\mathcal{S}, \mathcal{S}} & \mathbf{\Sigma}_{\mathcal{S}, \mathcal{L}} \\ \mathbf{\Sigma}_{\mathcal{L}, \mathcal{S}} & \mathbf{\Sigma}_{\mathcal{L}, \mathcal{L}} \end{pmatrix}$ where $\mathbf{\Sigma}_{\mathcal{S}, \mathcal{S}}$ is the submatrix of $\mathbf{\Sigma}$ corresponding to \mathcal{S} , we have

$$\frac{|\mathbf{\Sigma}_{\mathcal{S} \cup \mathcal{L}, \mathcal{S} \cup \mathcal{L}}|}{|\mathbf{\Sigma}_{\mathcal{L}, \mathcal{L}}|} = \frac{|\mathbf{\Sigma}_{\mathcal{L}, \mathcal{L}}| |\mathbf{\Sigma}_{\mathcal{S}, \mathcal{S}} - \mathbf{\Sigma}_{\mathcal{S}, \mathcal{L}} \mathbf{\Sigma}_{\mathcal{L}, \mathcal{L}}^{-1} \mathbf{\Sigma}_{\mathcal{L}, \mathcal{S}}|}{|\mathbf{\Sigma}_{\mathcal{L}, \mathcal{L}}|} = |\mathbf{\Sigma}_{\mathcal{S}|\mathcal{L}}|,$$

where $\Sigma_{S|\mathcal{L}}$ is the covariance matrix of the conditional distribution $p(\mathbf{y}_S|\mathbf{y}_{\mathcal{L}})$ and is naturally positive semidefinite. Therefore, the problem (1) is equivalent to

$$\min_{S \subseteq \mathcal{U}} \text{tr}(\Sigma_{S,S} - \Sigma_{S,\mathcal{L}} \Sigma_{\mathcal{L},\mathcal{L}}^{-1} \Sigma_{\mathcal{L},S}). \quad (2)$$

Discriminability. The tendency of an image \mathbf{x}_i belonging to a class \mathcal{C}_j is modeled by the average commute time between \mathbf{x}_i and all the images in \mathcal{C}_j , which is formally represented by

$$\bar{T}(\mathbf{x}_i, \mathcal{C}_j) = \frac{1}{n_{\mathcal{C}_j}} \sum_{\mathbf{x}_{i'} \in \mathcal{C}_j} T(\mathbf{x}_i, \mathbf{x}_{i'}). \quad (3)$$

In Eq. (3), $n_{\mathcal{C}_j}$ denotes the number of images in the class \mathcal{C}_j ; $T(\mathbf{x}_i, \mathbf{x}_{i'})$ is the commute time [41] between images \mathbf{x}_i and $\mathbf{x}_{i'}$, which is defined by

$$T(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{k=1}^n h(\lambda_k) (u_{ki} - u_{ki'})^2,$$

where $0 = \lambda_1 \leq \dots \leq \lambda_n$ are the eigenvalues of \mathbf{L} , and $\mathbf{u}_1, \dots, \mathbf{u}_n$ are the associated eigenvectors; u_{ki} denotes the i -th element of \mathbf{u}_k ; $h(\lambda_k) = 1/\lambda_k$ if $\lambda_k \neq 0$ and $h(\lambda_k) = 0$ otherwise.

Therefore, suppose \mathcal{C}_1 and \mathcal{C}_2 are the two closest classes to $\mathbf{x}_i \in \mathcal{U}$ measured by average commute time, then \mathbf{x}_i is discriminable if the gap $M(\mathbf{x}_i) = \bar{T}(\mathbf{x}_i, \mathcal{C}_2) - \bar{T}(\mathbf{x}_i, \mathcal{C}_1)$ is large. That is, the simplest curriculum in view of discriminability is found by solving

$$\min_{S=\{\mathbf{x}_{i_k} \in \mathcal{U}\}_{k=1}^s} \sum_{k=1}^s 1/M(\mathbf{x}_{i_k}), \quad (4)$$

where s is the amount of images in the set S .

By combining Eqs. (2) and (4), we arrive at the following optimization problem:

$$\min_{S=\{\mathbf{x}_{i_k} \in \mathcal{U}\}_{k=1}^s} \text{tr}(\Sigma_{S,S} - \Sigma_{S,\mathcal{L}} \Sigma_{\mathcal{L},\mathcal{L}}^{-1} \Sigma_{\mathcal{L},S}) + \sum_{k=1}^s 1/M(\mathbf{x}_{i_k}). \quad (5)$$

In each propagation, the seed labels will be propagated to the unlabeled images that are direct neighbors (denoted by the neighbouring set \mathcal{B}) of \mathcal{L} on graph \mathcal{G} , so we only need to choose s distinct images from \mathcal{B} . Therefore, we introduce a binary selection matrix $\mathbf{S} \in \{1, 0\}^{b \times s}$ (b is the size of the set \mathcal{B}) such that $\mathbf{S}^\top \mathbf{S} = \mathbf{I}_{s \times s}$. The element $\mathbf{S}_{ij} = 1$ means that the i -th image in \mathcal{B} is selected as the j -th element in the curriculum S . The orthogonality constraint $\mathbf{S}^\top \mathbf{S} = \mathbf{I}_{s \times s}$ imposed on \mathbf{S} ensures that every image is selected only once in S . The problem (5) is subsequently reformulated to the following matrix form:

$$\begin{aligned} \min_{\mathbf{S}} & \text{tr}(\mathbf{S}^\top \Sigma_{\mathcal{B},\mathcal{B}} \mathbf{S} - \mathbf{S}^\top \Sigma_{\mathcal{B},\mathcal{L}} \Sigma_{\mathcal{L},\mathcal{L}}^{-1} \Sigma_{\mathcal{L},\mathcal{B}} \mathbf{S}) \\ & + \text{tr}(\mathbf{S}^\top \mathbf{M} \mathbf{S}), \\ \text{s.t. } & \mathbf{S} \in \{1, 0\}^{b \times s}, \mathbf{S}^\top \mathbf{S} = \mathbf{I}_{s \times s}, \end{aligned} \quad (6)$$

where $\mathbf{M} \in \mathbb{R}^{b \times b}$ is a diagonal matrix with the diagonal elements $\mathbf{M}_{ii} = 1/M(\mathbf{x}_i)$ for any $\mathbf{x}_i \in \mathcal{B}$. By denoting $\mathbf{R} = \Sigma_{\mathcal{B},\mathcal{B}} - \Sigma_{\mathcal{B},\mathcal{L}} \Sigma_{\mathcal{L},\mathcal{L}}^{-1} \Sigma_{\mathcal{L},\mathcal{B}} + \mathbf{M}$, the curriculum selection

model for single-modal case is simplified as

$$\begin{aligned} \min_{\mathbf{S}} & \text{tr}(\mathbf{S}^\top \mathbf{R} \mathbf{S}), \\ \text{s.t. } & \mathbf{S} \in \{1, 0\}^{b \times s}, \mathbf{S}^\top \mathbf{S} = \mathbf{I}_{s \times s}. \end{aligned} \quad (7)$$

B. Multi-modal Curriculum Generation

Single-modal curriculums cannot always render satisfactory performance (demonstrated in Section IV-A), we therefore extend the single-modal model (7) to multi-modal cases. The high level idea is to force the V teachers to reach a consensus on selecting the optimal curriculum \mathbf{S}^* . This is formulated as an optimization problem, which can be solved by relaxing the binary selection matrices to continuous ones, and then conducting the standard alternating minimization. Each subproblem in the alternating minimization is constrained by an orthogonal constraint, and can be optimized by the existing solver on matrix manifold.

In order to regulate the selection matrices $\mathbf{S}^{(v)}$ ($v = 1, \dots, V$) generated by V teachers to compromise to a common \mathbf{S}^* , we define the following optimization:

$$\begin{aligned} \min_{\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{S}^*} & \sum_{v=1}^V \text{tr}(\mathbf{S}^{(v)\top} \mathbf{R}^{(v)} \mathbf{S}^{(v)}) + \beta \sum_{v=1}^V \left\| \mathbf{S}^{(v)} - \mathbf{S}^* \right\|_F^2, \\ \text{s.t. } & \mathbf{S}^* \in \{1, 0\}^{b \times s}, \mathbf{S}^{*\top} \mathbf{S}^* = \mathbf{I}_{s \times s}, \\ & \mathbf{S}^{(v)} \in \{1, 0\}^{b \times s}, \mathbf{S}^{(v)\top} \mathbf{S}^{(v)} = \mathbf{I}_{s \times s}, \text{ for } v=1, \dots, V. \end{aligned} \quad (8)$$

where the first term in the objective function shares the similar purpose with the objective in Eq. (7), which requires all teachers to select the simplest images according to their modality viewpoints. The second term makes the teachers maximally agree with each other and produce the consistent curriculum, where “ $\|\cdot\|_F$ ” computes the Frobenius norm. $\beta > 0$ is the trade-off parameter. However, The binary constraints turn the optimization (8) into an integer programming which is generally NP-hard. To make problem (8) tractable, we relax the discrete constraints to continuous nonnegative constraints and achieve the following expression:

$$\begin{aligned} \min_{\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{S}^*} & \sum_{v=1}^V \text{tr}(\mathbf{S}^{(v)\top} \mathbf{R}^{(v)} \mathbf{S}^{(v)}) + \beta \sum_{v=1}^V \left\| \mathbf{S}^{(v)} - \mathbf{S}^* \right\|_F^2, \\ \text{s.t. } & \mathbf{S}^* \geq \mathbf{O}_{b \times s}, \mathbf{S}^{*\top} \mathbf{S}^* = \mathbf{I}_{s \times s}, \\ & \mathbf{S}^{(v)} \geq \mathbf{O}_{b \times s}, \mathbf{S}^{(v)\top} \mathbf{S}^{(v)} = \mathbf{I}_{s \times s}, \text{ for } v=1, \dots, V \end{aligned} \quad (9)$$

where \mathbf{O} denotes the zero matrix. A local minimizer of problem (9) can be obtained by alternatively optimizing $\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}$, and \mathbf{S}^* with other variables remaining fixed.

Updating $\mathbf{S}^{(v)}$. To obtain the optimal selection matrix $\mathbf{S}^{(v)}$ where v takes a value from $1, \dots, V$, we treat \mathbf{S}^* and $\mathbf{S}^{(v')}$ ($v' \neq v$) as constant variables, and then the $\mathbf{S}^{(v)}$ -subproblem is derived as

$$\begin{aligned} \min_{\mathbf{S}^{(v)}} & \text{tr}(\mathbf{S}^{(v)\top} \mathbf{R}^{(v)} \mathbf{S}^{(v)}) + \beta \left\| \mathbf{S}^{(v)} - \mathbf{S}^* \right\|_F^2 \\ \text{s.t. } & \mathbf{S}^{(v)} \geq \mathbf{O}_{b \times s}, \mathbf{S}^{(v)\top} \mathbf{S}^{(v)} = \mathbf{I}_{s \times s}. \end{aligned} \quad (10)$$

The nonnegative constraint in the above optimization can be easily tackled by the method of augmented Lagrangian multiplier (ALM). The main idea of ALM is to transform a constrained optimization problem into a non-constrained problem by incorporating penalty terms. Compared with the traditional Lagrangian method, ALM adds an additional quadratic penalty function to the objective, which leads to faster convergence rate and lower computational cost [42]. However, in our case the orthogonal constraint cannot be directly degenerated into the augmented Lagrangian function because it defines a nonconvex feasible region on the Stiefel manifold (the Stiefel manifold is the set of all $m_1 \times m_2$ matrices satisfying the orthogonal constraint, *i.e.* $St(m_1, m_2) = \{\mathbf{X} \in \mathbb{R}^{m_1 \times m_2} : \mathbf{X}^\top \mathbf{X} = \mathbf{I}_{m_2 \times m_2}\}$). As a result, we only incorporate the nonnegative constraint into the augmented Lagrangian function and project the result produced by gradient descent back onto the Stiefel manifold in each iteration. The augmented Lagrangian function is

$$\begin{aligned} L(\mathbf{S}^{(v)}, \mathbf{\Lambda}^{(v)}, \mathbf{T}^{(v)}, \sigma^{(v)}) \\ = \text{tr}(\mathbf{S}^{(v)\top} \mathbf{R}^{(v)} \mathbf{S}^{(v)}) + \beta \left\| \mathbf{S}^{(v)} - \mathbf{S}^* \right\|_{\text{F}}^2 \\ + \text{tr}(\mathbf{\Lambda}^{(v)\top} (\mathbf{S}^{(v)} - \mathbf{T}^{(v)})) + \frac{\sigma^{(v)}}{2} \left\| \mathbf{S}^{(v)} - \mathbf{T}^{(v)} \right\|_{\text{F}}^2, \end{aligned}$$

where $\mathbf{\Lambda}^{(v)} \in \mathbb{R}^{b \times s}$ is the Lagrangian multiplier, $\mathbf{T}^{(v)} \in \mathbb{R}^{b \times s}$ is a nonnegative auxiliary matrix, and $\sigma^{(v)} > 0$ is the penalty coefficient. Therefore, $\mathbf{S}^{(v)}$ is updated by

$$\mathbf{S}^{(v)} := \text{Proj}_{St} \left[\mathbf{S}^{(v)} - \tau \nabla_{\mathbf{S}^{(v)}} L(\mathbf{S}^{(v)}, \mathbf{\Lambda}^{(v)}, \mathbf{T}^{(v)}, \sigma^{(v)}) \right], \quad (11)$$

where $\nabla_{\mathbf{S}^{(v)}} L(\mathbf{S}^{(v)}, \mathbf{\Lambda}^{(v)}, \mathbf{T}^{(v)}, \sigma^{(v)})$ computes the gradient of the augmented Lagrangian function L on $\mathbf{S}^{(v)}$, and τ is the step size chosen by the backtracking line search method. The projection $\text{Proj}_{St}[\mathbf{X}]$ has a closed form based on the unitary factor of a polar decomposition on \mathbf{X} (See Proposition 7 in [43]).

The auxiliary matrix $\mathbf{T}^{(v)}$ is updated by the conventional rule in the augmented Lagrangian method, which is $\mathbf{T}_{ij}^{(v)} := \max(0, \mathbf{S}_{ij}^{(v)} + \mathbf{\Lambda}_{ij}^{(v)} / \sigma^{(v)})$. The above iterative process for solving the subproblem (10) is summarized in Algorithm 1, and is guaranteed to be convergent [44].

Updating \mathbf{S}^* . The \mathbf{S}^* -subproblem is formulated as

$$\begin{aligned} \min_{\mathbf{S}^*} \quad & \sum_{v=1}^V \left\| \mathbf{S}^{(v)} - \mathbf{S}^* \right\|_{\text{F}}^2, \\ \text{s.t.} \quad & \mathbf{S}^* \geq \mathbf{O}_{b \times s}, \quad \mathbf{S}^{*\top} \mathbf{S}^* = \mathbf{I}_{s \times s} \end{aligned} \quad (12)$$

which can be solved via the same way as the $\mathbf{S}^{(v)}$ -subproblem. Therefore, we omit the explanation for optimizing Eq. (12).

The adopted alternating minimization between $\mathbf{S}^{(v)}$ and \mathbf{S}^* ensures that the objective function value of Eq. (9) always decreases. Besides, this objective function is lower bounded by 0 since the matrices $\mathbf{R}^{(v)}$ ($v = 1, \dots, V$) are positive definite. Therefore, the entire alternating optimization process is guaranteed to converge, and the obtained \mathbf{S}^* is agreed on by all the teachers. However, the solution \mathbf{S}^* for Eq. (9) is continuous, which does not satisfy the original binary

Algorithm 1 The algorithm for solving $\mathbf{S}^{(v)}$ -subproblem (10)

```

1: Input:  $\mathbf{R}^{(v)}, \mathbf{S}^*, \mathbf{S}^{(v)} \in St, \mathbf{\Lambda}^{(v)} = \mathbf{O}, \sigma^{(v)} = 1, \rho = 1.2, \beta,$ 
    $iter = 0$ 
2: repeat
3:   // Compute  $\mathbf{T}^{(v)}$ 
4:    $\mathbf{T}_{ij}^{(v)} = \max(0, \mathbf{S}_{ij}^{(v)} + \mathbf{\Lambda}_{ij}^{(v)} / \sigma^{(v)});$ 
5:   // Update  $\mathbf{S}^{(v)}$  by using Eq. (11)
6:    $\mathbf{S}^{(v)} := \text{Proj}_{St} \left[ \mathbf{S}^{(v)} - \tau \nabla_{\mathbf{S}^{(v)}} L(\mathbf{S}^{(v)}, \mathbf{\Lambda}^{(v)}, \mathbf{T}^{(v)}, \sigma^{(v)}) \right];$ 
7:   // Update variables
8:    $\mathbf{\Lambda}_{ij}^{(v)} := \max(0, \mathbf{\Lambda}_{ij}^{(v)} - \sigma^{(v)} \mathbf{S}_{ij}^{(v)});$ 
9:    $\sigma^{(v)} := \min(\rho \sigma^{(v)}, 10^{10}); iter := iter + 1;$ 
10: until Convergence
11: Output:  $\mathbf{S}^{(v)}$  that minimizes Eq. (10)

```

constraint in problem (8). Consequently, we then discretize \mathbf{S}^* to binary values via a simple greedy procedure. In detail, we find the largest element in \mathbf{S}^* , and record its row and column; then from the unrecorded columns and rows we search the largest element and mark it again. This procedure is repeated until s elements have been found. The rows of these s elements indicate the simplest images selected for propagation.

C. Multi-modal Classification with Feedback

When the overall optimal curriculum $\mathcal{S}^* = \{\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_s^*\}$ is specified by the teachers, the learner will propagate the labels from \mathcal{L} to these s images from each of the V modalities. The output label matrices $\mathbf{F}^{(1)}, \dots, \mathbf{F}^{(V)}$ are then fused into a consistent \mathbf{F} . We employ the label propagation algorithm [11] as the learner because it is naturally incremental and does not require retraining with the arrival of a new curriculum. Under the t -th propagation, the iterative model for a specific modality v is:

$$\mathbf{F}_i^{(v)[t]} = \begin{cases} \mathbf{P}_i^{(v)} \mathbf{F}^{[t-1]}, & \mathbf{x}_i \in (\mathcal{S}^{*[1]} \cup \dots \cup \mathcal{S}^{*[t-1]}) \cup \mathcal{S}^{*[t]} \\ \mathbf{F}_i^{[0]}, & \mathbf{x}_i \in \mathcal{L}^{[0]} \cup (\mathcal{U}^{[0]} - \mathcal{S}^{*[1]} \cup \dots \cup \mathcal{S}^{*[t]}) \end{cases} \quad (13)$$

where $\mathbf{F}_i^{(v)[t]}$ denotes the i -th row of the matrix $\mathbf{F}^{(v)[t]}$, $\mathbf{F}^{[t-1]}$ is the consistent label matrix produced by the previous propagation, $\mathbf{P}_i^{(v)}$ represents the i -th row of the transition matrix $\mathbf{P}^{(v)}$ calculated by $\mathbf{P}^{(v)} = \mathbf{D}^{(v)-1} \mathbf{W}^{(v)}$, and $\mathcal{U}^{[0]} - \mathcal{S}^{*[1]} \cup \dots \cup \mathcal{S}^{*[t]}$ is the complementary set of $\mathcal{S}^{*[1]} \cup \dots \cup \mathcal{S}^{*[t]}$ in $\mathcal{U}^{[0]}$. The superscript $[t]$ represents the t -th propagation. Eq. (13) suggests that the labels of the t -th curriculum and previously “learned” images will change during the t -th propagation, whereas the labels of the initially labeled images in $\mathcal{L}^{[0]}$ and the unclassified unlabeled images in $\mathcal{U}^{[0]} - \mathcal{S}^{*[1]} \cup \dots \cup \mathcal{S}^{*[t]}$ are kept unchanged, as also suggested by Zhu *et al.* [11]. The initial state for \mathbf{x}_i ’s label vector $\mathbf{F}_i^{[0]}$ is

$$\mathbf{F}_i^{[0]} := \begin{cases} (1/c, \dots, 1/c), & \mathbf{x}_i \in \mathcal{U}^{[0]} \\ \left(0, \dots, \underset{j\text{-th element}}{1}, \dots, 0 \right), & \mathbf{x}_i \in \mathcal{C}_j \in \mathcal{L}^{(0)} \end{cases}, \quad (14)$$

where c is the total number of classes. The formulations of Eqs. (13) and (14) maintain the probability interpretation $\sum_{j=1}^c \mathbf{F}_{ij}^{[t]} = 1$ for any image \mathbf{x}_i and all t -th ($t = 0, 1, 2, \dots$)

propagations. Consequently, the integrated label matrix $\mathbf{F}^{[t]}$ is computed by:

$$\mathbf{F}^{[t]} = \sum_{v=1}^V \omega^{(v)[t]} \mathbf{F}^{(v)[t]}, \quad (15)$$

where the weights are

$$\omega^{(v)[t]} = \frac{\exp\left(-\|\mathbf{S}^{(v)[t]} - \mathbf{S}^{*[t]}\|_{\mathbf{F}}^2\right)}{\sum_{v=1}^V \exp\left(-\|\mathbf{S}^{(v)[t]} - \mathbf{S}^{*[t]}\|_{\mathbf{F}}^2\right)}. \quad (16)$$

Eq. (16) imposes a large weight on the v -th label matrix $\mathbf{F}^{(v)[t]}$ in Eq. (15) if the corresponding teacher generates a similar curriculum to the overall optimal curriculum $\mathbf{S}^{*[t]}$. This is because small $\|\mathbf{S}^{(v)[t]} - \mathbf{S}^{*[t]}\|_{\mathbf{F}}^2$ suggests that, compared to other modality viewpoints, the $\mathbf{S}^{*[t]}$ agrees more on the simplest curriculum generated from the v -th modality, thus its propagation result should be emphasized.

When the t -th learning has been completed, the learner will deliver an overall feedback to the teachers to assist them in designing the suitable $(t+1)$ -th curriculum. Intuitively, if the classification result is confident, the teachers may assign a ‘‘heavier’’ curriculum to the learner in the $(t+1)$ -th curriculum; that is, the size of $\mathbf{S}^{*[t+1]}$ (*i.e.* $s^{[t+1]}$) can be increased. For example, suppose we have $c = 3$ classes in total, then for a single image \mathbf{x}_i , its classification result is confident if it has a label vector $\mathbf{F}_i = [1, 0, 0]$, $[0, 1, 0]$, or $[0, 0, 1]$, which means that \mathbf{x}_i definitely belongs to the class 1, 2 or 3, respectively. In contrast, if \mathbf{x}_i 's label vector is $\mathbf{F}_i = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$, its learning result is not satisfactory because $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ cannot provide any cue for determining its class. Such learning confidence is evaluated by the entropy of $\mathbf{S}^{*[t]}$'s label matrix $\mathbf{F}_{\mathbf{S}^{*[t]}}$ (*i.e.* $H(\mathbf{F}_{\mathbf{S}^{*[t]}})$) [18], which is formally defined by

$$\begin{aligned} \text{Conf}(\mathbf{F}_{\mathbf{S}^{*[t]}}) &= \exp\left[-\gamma^{[t]} \frac{1}{s^{[t]}} H(\mathbf{F}_{\mathbf{S}^{*[t]}})\right] \\ &= \exp\left[\frac{\gamma^{[t]}}{s^{[t]}} \sum_{i=1}^{s^{[t]}} \sum_{j=1}^c (\mathbf{F}_{\mathbf{S}^{*[t]}})_{ij} \log_c (\mathbf{F}_{\mathbf{S}^{*[t]}})_{ij}\right], \end{aligned} \quad (17)$$

where $\gamma^{[t]}$ controls the learning rate and is gradually decreased by $\gamma^{[t]} = \gamma^{[t-1]}/\eta$ ($\eta > 1$) so that more images will be incorporated by the curriculums in later propagations. This manipulation is reasonable because the rich knowledge accumulated in previous propagations later helps to boost the learning speed. It is easy to verify that $\text{Conf}(\mathbf{F}_{\mathbf{S}^{*[t]}}) \in (0, 1]$, and $\text{Conf}(\mathbf{F}_{\mathbf{S}^{*[t]}})$ touches its maximum value 1 if every row in $\mathbf{F}_{\mathbf{S}^{*[t]}}$ is a $\{0, 1\}$ -binary vector with only one 1. This suggests that the class labels of all propagated images are clearly indicated. In contrast, if all the elements in $\mathbf{F}_{\mathbf{S}^{*[t]}}$ are close to the ambiguous value $1/c$, $\text{Conf}(\mathbf{F}_{\mathbf{S}^{*[t]}})$ will obtain a very small value. Based on Eq. (17), the number of simplest images in the $(t+1)$ -th curriculum is

$$s^{[t+1]} = \left\lceil b^{[t+1]} \cdot \text{Conf}(\mathbf{F}_{\mathbf{S}^{*[t]}}) \right\rceil, \quad (18)$$

where $b^{[t+1]}$ is the size of neighbouring set $\mathcal{B}^{[t+1]}$ in the $(t+1)$ -th propagation, and $\lceil \cdot \rceil$ rounds up the element to the nearest integer.

The above teaching-then-learning process iterates until all

Algorithm 2 MMCL for semi-supervised image classification

```

1: Input:  $l$  labeled images  $\mathcal{L} = \{\mathbf{x}_1, \dots, \mathbf{x}_l\}$  with known labels
 $y_1, \dots, y_l$  expressed in  $V$  modalities;  $u$  unlabeled images  $\mathcal{U} =$ 
 $\{\mathbf{x}_{l+1}, \dots, \mathbf{x}_{l+u}\}$  with unknown labels  $y_{l+1}, \dots, y_{l+u}$ ; Parameters
 $\beta, \gamma, \eta, \theta, \kappa$ ;
2: // Pre-processing
3:  $\forall v=1, \dots, V$ , compute  $\mathbf{W}^{(v)}$ ,  $\Sigma^{(v)}$  and  $\mathbf{L}^{(v)}$  corresponding to
 $V$  graphs  $\mathcal{G}^{(1)}, \dots, \mathcal{G}^{(V)}$ ;
4: // Multi-modal curriculum generation and learning
5: repeat
6: // Compute optimal curriculum  $\mathbf{S}^*$  by solving Eq. (9);
7: repeat
8: Update  $\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}$  sequentially by solving Eq. (10);
9: Update  $\mathbf{S}^*$  by solving Eq. (12);
10: until Convergence
11: // Conduct label propagation on each modality viewpoint
12: Compute the label matrix  $\mathbf{F}^{(v)}$  via Eq. (13);
13: Fuse  $V$  label matrices to  $\mathbf{F}$  via Eq. (15);
14: // Establish learning feedback
15: Compute  $\text{Conf}(\mathbf{F}_{\mathbf{S}^{*[t]}})$  via Eq. (17);
16: Compute the size of  $(t+1)$ -th curriculum via Eq. (18);
17: // Update sets
18:  $\mathcal{L} := \mathcal{L} \cup \mathbf{S}^*$ ;  $\mathcal{U} := \mathcal{U} - \mathbf{S}^*$ ;  $\gamma := \gamma/\eta$ ;
19: until  $\mathcal{U} = \emptyset$ ;
20: Compute the steady state  $\bar{\mathbf{F}}^{*(v)}$  on each graph via Eq. (20);
21: Compute the final learned label matrix by  $\bar{\mathbf{F}}^* = \frac{1}{V} \sum_{v=1}^V \bar{\mathbf{F}}^{*(v)}$ ;
22: Assign labels to images via  $j = \arg \max_{j' \in \{1, \dots, c\}} \bar{\mathbf{F}}_{ij'}^*$ ;
23: Output: Class labels  $y_{l+1}, \dots, y_{l+u}$ ;

```

the unlabeled images have been used, and the integrated label matrix thus obtained is denoted as $\bar{\mathbf{F}}$. We then start from $\bar{\mathbf{F}}^{[0]} := \bar{\mathbf{F}}$ and use the iterative formula [45] to drive the propagations on each graph $\mathcal{G}^{(v)}$ to the steady state:

$$\bar{\mathbf{F}}^{(v)[t]} = \theta \mathbf{P}^{(v)} \bar{\mathbf{F}}^{(v)[t-1]} + (1 - \theta) \bar{\mathbf{F}}^{[0]}, \quad (19)$$

where the parameter $\theta > 0$ balances the labels propagated from other images, and $\bar{\mathbf{F}}^{[0]}$ that is produced by the teaching-then-learning process. We set $\theta = 0.05$ to ensure the final result will be maximally consistent with the labels produced by multi-modal curriculum learning. Eq. (19) is proved [11] to converge to

$$\bar{\mathbf{F}}^{*(v)} = \lim_{t \rightarrow \infty} \bar{\mathbf{F}}^{(v)[t]} = (1 - \theta)(\mathbf{I} - \theta \mathbf{P}^{(v)})^{-1} \bar{\mathbf{F}}^{[0]}, \quad (20)$$

and the final learned label matrix is $\bar{\mathbf{F}}^* = \frac{1}{V} \sum_{v=1}^V \bar{\mathbf{F}}^{*(v)}$. Eventually, the image \mathbf{x}_i is assigned to the j -th class, which satisfies $j = \arg \max_{j' \in \{1, \dots, c\}} \bar{\mathbf{F}}_{ij'}^*$. The complete MMCL algorithm for semi-supervised image classification is outlined in Algorithm 2.

IV. EXPERIMENTAL RESULTS

In this section, we first validate the motivation of our MMCL algorithm on a small database (Section IV-A) and then compare MMCL with several state-of-the-art methods on eight practical image datasets (Section IV-B). The MMCL parameters in all the experiments are set to $\beta = 10$, $\gamma = 3$ and $\eta = 1.1$, and the parametric sensitivity will be studied in Section IV-C.

In this section, all the images in the adopted datasets are represented by the 72-dimensional Pyramid Histogram Of Gradients (PHOG) [46], 512-dimensional GIST [47], and

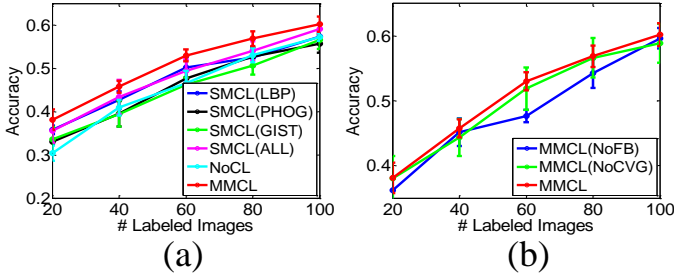


Fig. 2. Validations of our MMCL motivation. (a) compares the performance of multi-modal curriculum learning with single-modal curriculum learning and plain learning; (b) shows the effects of some key steps in our MMCL model.

256-dimensional Local Binary Patterns (LBP) [48] features. Therefore, three different modalities are formed to serve as the input of all the compared methods. Note that these feature descriptors are histogram-based and every element in a feature vector falls into $[0, 1]$, so none of them will dominate the learning performance.

A. Algorithm Validation

Firstly, we use a subset of the *Caltech256* dataset [49] to demonstrate two arguments: 1) curriculum learning is critical to improving classification performance, and 2) MMCL is superior to single-modal curriculum learning (SMCL). The images of dog, goose, swan and zebra in *Caltech256* are employed, and each type of animals contains 80 examples. For MMCL, each feature constitutes a modality, and Algorithm 2 is utilized to accomplish the classification. We also report the results of SMCL, of which the model explained in Section III-A is adopted to handle each of the three feature descriptors (denoted “SMCL(LBP)”, “SMCL(PHOG)”, and “SMCL(GIST)”, respectively). Furthermore, we concatenate these three different feature vectors to a long vector, and apply SMCL again to simultaneously utilize the three feature modalities (denoted “SMCL(ALL)”). At last, the plain learning model [11] in our algorithm is implemented based on the concatenated long feature vectors, to test the performance without curriculum learning (denoted “NoCL”).

We evaluate the classification accuracies of tested methods under different sizes of labeled set, and the experiment under each size is conducted ten times with different initially labeled images. The reported accuracies are then obtained by averaging over the outputs of these ten independent runs. Fig. 2(a) presents the result, in which the record under each number of labeled images includes mean accuracy as well as the standard deviation. It can be observed that MMCL and SMCL(ALL) always outperform NoCL under different numbers of labeled images, therefore the argument 1) above is verified. Specifically, SMCL(ALL) and MMCL outperform NoCL with the margins about 1%~5% and 3%~8%, respectively, therefore the effectiveness of curriculum learning has been demonstrated. Moreover, we see that MMCL achieves the highest record compared to all the other single-modal counterparts, which explicitly justifies the argument 2). It is also clearly shown that the concatenation of all different features (SMCL(ALL)) generally yields better performance

than simply working on a single feature (e.g. SMCL(PHOG), SMCL(LBP) and SMCL(GIST)). Therefore, properly exploiting different modalities for curriculum learning is superior to simply working on single modality. The reason lies in that multiple feature modalities convey richer image information than the single modality, and this is also consistent with our general understanding.

Next, we demonstrate the effectiveness of a number of key steps in our MMCL model, such as the establishment of learning feedback Eq. (17) and the convergence of propagations Eq. (20). To demonstrate the contribution of learning feedback, we remove the feedback and fix the number of selected simplest images in each propagation t to $\min(20, b^{[t]})$ to generate the accuracy (see “MMCL(NoFB)” in Fig. 2(b)). To show the importance of converge, we plot the accuracy generated by the non-convergent \bar{F} (see “MMCL(NoCVG)”). By comparing the three curves in Fig. 2(b), we clearly see that performance decreases in the absence of each of the two manipulations. Therefore, incorporating these steps in our model contributes to improved accuracy.

Lastly, we visualize the curriculum images selected by our MMCL during the entire teaching and learning process. When the number of labeled images is 60, our MMCL takes totally 14 propagations to classify all the unlabeled images, and the selected simplest images under different iterations t are provided in Fig. 3. We can see that during the initial stage of the propagation, i.e. $t = 1 \sim 2$, the teachers in MMCL tend to select the images containing complete objects with regular appearances. Besides, the backgrounds in these images are also generally clean and are very different from the foreground objects. When $t = 8 \sim 9$, we see that some of the selected images only contain part of the objects (e.g. dog, goose and zebra), or reflect the objects with abnormal behavior compared to their normal conditions (e.g. swan). During the final stage of propagations, i.e. $t = 11 \sim 14$, the curriculum examples are quite difficult because of the multiple crowded objects (e.g. dog, goose and zebra) or the uncommon observation angle (e.g. swan). Therefore, the introduced teachers in MMCL can accurately evaluate the difficulty level of every unlabeled image, and effectively organize the entire propagation process so that all the images are classified from simple to difficult.

B. Comparison with Other Methods

To further demonstrate the strength of the proposed method, we compare MMCL with other state-of-the-art algorithms on some typical image datasets.

Datasets. Eight image datasets with different contents are adopted for our experiments: *CaltechAnimal* [49] for animal classification, *Architecture* [50] for architecture style recognition, *MSRC*¹ for natural image classification, *UIUC* [51] for sports event recognition, *Scene15* [52] for scene categorization, *ORLFace*² for face recognition, and *CIFAR100* [53] and *NUS-WIDE* [54] for general image classification. Of these, *CaltechAnimal* is a subset of *Caltech256* consisted of nine different animals, and *NUS-WIDE* is formed by only

¹<http://research.microsoft.com/en-us/projects/objectclassrecognition/>

²<http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>



Fig. 3. The selected simplest images during the entire propagation process. It can be observed that the difficulty level of curriculum images gradually increases with the propagation proceeds.

preserving the classes that have more than 100 images in the original dataset. The details of these datasets are summarized in Table I. Besides, some example images of these datasets are provide in Fig. 5, which reflects that accurately classifying the images in these datasets is a very challenging task.

Baselines. Five semi-supervised image classifiers are adopted for comparison. The Gaussian Field and Harmonic Functions (GFHF) [11] is a classical SSL algorithm, which also serves as the learner in our proposed methodology. Dynamic Label Propagation (DLP) [8] is a recently proposed single-modal semi-supervised image classifier. For multi-modal baselines, Adaptive Multi-Modal Semi-Supervised classifier (AMMSS) [1] and Sparse Multiple Graph Integration (SMGI) [55] are employed for comparison as they also operate on multiple graphs. The results of SMCL introduced in Section III-A are also reported. Among the compared existing methods, the codes of AMMSS and SMGI are directly provided by the authors. We implement GFHF and DLP by ourselves because both methods can be easily reproduced in lines of MATLAB code.

As explained at the beginning of Section IV, GIST, LBP and PHOG features constitute three modalities for the multi-modal algorithms such as AMMSS, SMGI and our MMCL. These three descriptors are concatenated into a long feature vector for single-modal methodologies including GFHF, DLP and SMCL.

Experimental settings. Similar to Section IV-A, the accuracies of all the algorithms are evaluated under different selections of initially labeled images, and at least one labeled image is selected in each class. The reported accuracies and standard deviations are calculated as the mean value of the outputs of ten independent runs.

To achieve fair comparison, the identical 10-NN graphs are built via adaptive edge weighting [38] for all the methods except DLP. In DLP, the 10-NN graph is built by leveraging the Gaussian kernel as required. The key parameters in SMGI are optimally tuned to $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$ via searching the grid $\{0.01, 0.1, 1, 10\}$, and r and λ in AMMSS are respectively set to 0.5 and 10. As recommended by the authors, we adjust α and λ in DLP to 0.05 and 0.1 throughout the experiments.

Results & Analyses. The experimental results on the eight datasets are shown in Fig. 4. Figs. 4 (a)~(h) indicate that

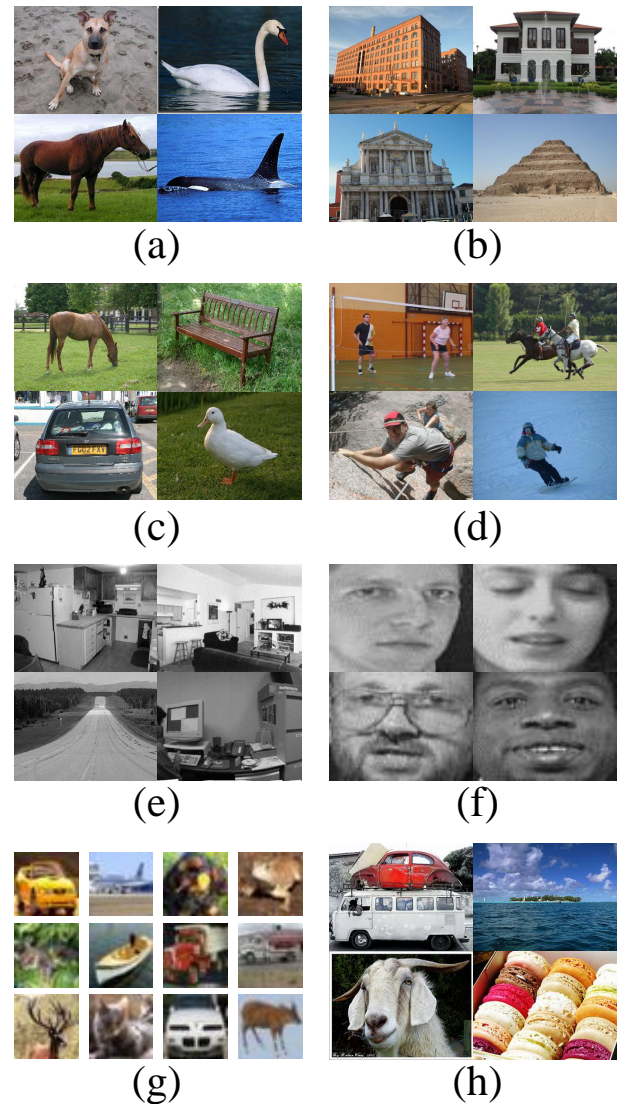


Fig. 5. Some example images in the adopted datasets. (a) is *CaltechAnimal* dataset, (b) is *Architecture* dataset, (c) is *MSRC* dataset, (d) is *UIUC* dataset, (e) is *Scene15* dataset, (f) is *ORLFace* dataset, (g) is *CIFAR100* dataset, and (h) is *NUS-WIDE* dataset.

TABLE I
AN OVERVIEW OF ADOPTED DATASETS.

	<i>CaltechAnimal</i>	<i>Architecture</i>	<i>MSRC</i>	<i>UIUC</i>	<i>Scene15</i>	<i>ORLFace</i>	<i>CIFAR100</i>	<i>NUS-WIDE</i>
# classes	9	25	20	8	15	40	100	112
# images	720	1000	589	1579	4485	400	60000	47254

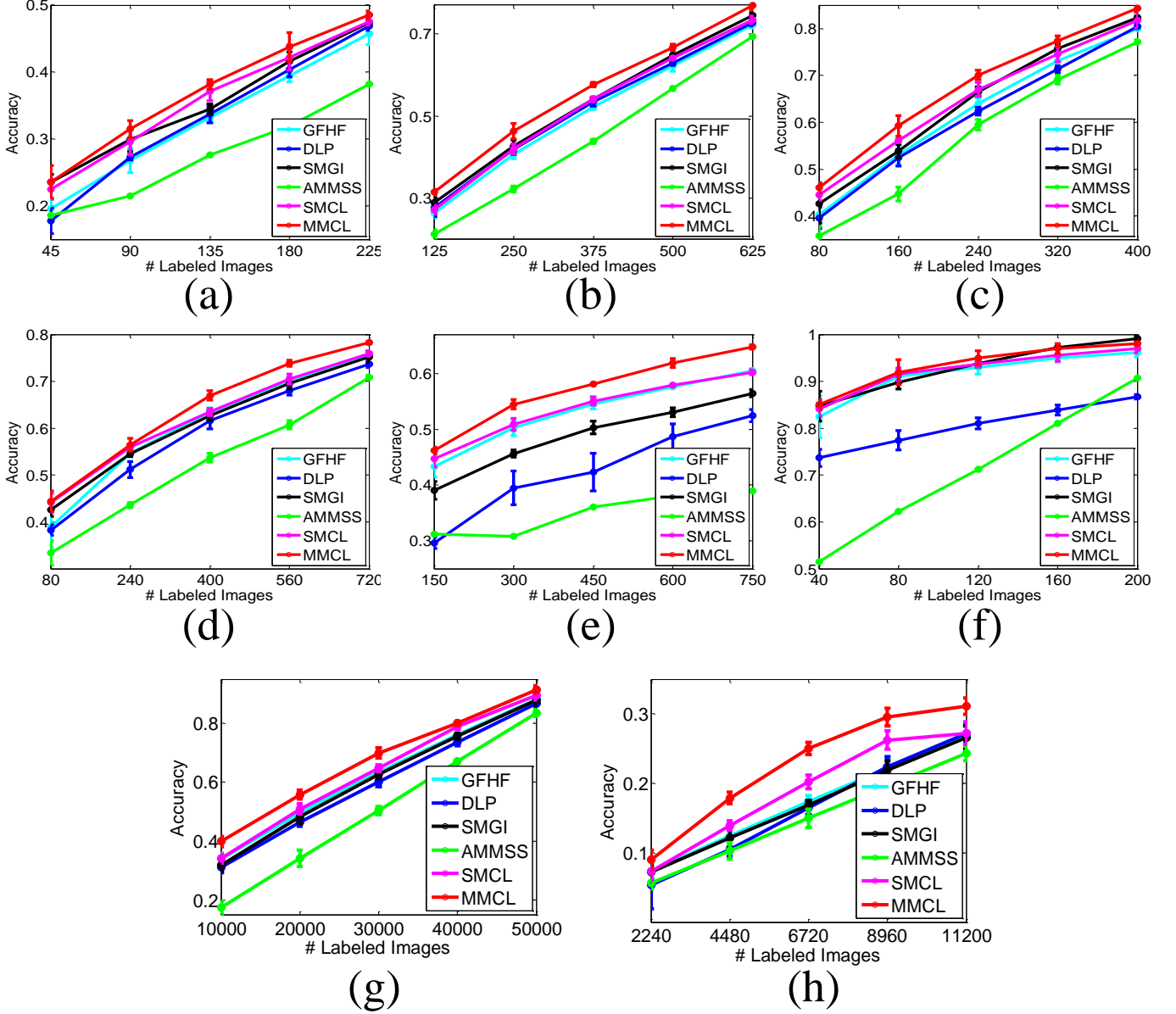


Fig. 4. Comparison of MMCL with other baselines on eight image datasets. (a) is *CaltechAnimal* dataset, (b) is *Architecture* dataset, (c) is *MSRC* dataset, (d) is *UIUC* dataset, (e) is *Scene15* dataset, (f) is *ORLFace* dataset, (g) is *CIFAR100* dataset, and (h) is *NUS-WIDE* dataset.

the proposed MMCL achieves the highest accuracy on the datasets *CaltechAnimal*, *Architecture*, *MSRC*, *UIUC*, *Scene15*, *CIFAR100* and *NUS-WIDE*. Numerically, MMCL leads SMCL with the margins approximately 2%, 3%, 3%, 3%, 4%, 4%, and 4% on the above seven datasets, respectively, and also significantly outperforms the best existing method SMGI or GFHF with the margins 3%, 3%, 4%, 4%, 4%, 5% and 5%, accordingly. On the *ORLFace* dataset revealed by Fig. 4

(f), MMCL performs better than SMGI when the number of labeled images ranges from 40 to 160. MMCL is slightly surpassed by SMGI when the size of labeled set is 200. Furthermore, we observe that on all the datasets the standard deviations of MMCL are very small, suggesting that MMCL is insensitive to the selection of initially labeled examples.

A number of other interesting facts can be observed from the experimental results. Firstly, SMCL (magenta curve) performs

	Bicycle	Chair	Croquet	Individual 28
GFHF	✗	✗	✗	✗
DLP	✗	✓	✗	✗
SMGI	✗	✗	✗	✓
AMMSS	✗	✗	✗	✗
SMCL	✗	✗	✓	✗
MMCL	✓	✓	✓	✓

Fig. 6. Classification results of the compared methods on several visually challenging images. The red crosses represent “incorrect classifications” while the green ticks denote “correct classifications”.

favorably compared to the other single-modal methods such as DLP (blue curve) and GFHF (cyan curve). MMCL (red curve) generally achieves better performance than other multi-modal approaches like SMGI (black curve) and AMMSS (green curve). Therefore, the established curriculums does help to optimize the learning process and generate encouraging classification performance. Secondly, sometimes the accuracy improvement brought by SMCL over GFHF is very marginal, such as on *Architecture*, *UIUC*, *Scene15*, *ORLFace*, and *CIFAR100* datasets. Comparatively, MMCL is significantly better than GFHF, and also enhances the performance of SMCL on all the datasets. Therefore, generating curriculums from multiple modalities is superior to only employing a single modality consisted of a long concatenated feature vector. This reflects that directly putting different types of features into a long vector is not an ideal way to handle multi-modal cases. The information from different sources should instead be integrated in an informed way, such that the strength of every modality can be fully exploited. These observations also comply with our findings in Section IV-A and again demonstrate the validity of the two arguments therein.

Further insight. As mentioned in the Introduction, MMCL can achieve higher classification due to its strong ability for handling difficult images. To illustrate this point, we investigate the classification correctness of the compared methods on some difficult images in the adopted datasets (see Fig. 6). In the “Bicycle” image, the occlusion and overlapping of bicycles make classification very difficult. The image belonging to “Chair” category contains one table and multiple chairs. The two men in the “Croquet” image are very small, and identifying their activities is nontrivial. The person in “Individual 28” wears a pair of glasses, which poses a great difficulty for accurate face recognition. Though these example images are visually challenging, MMCL is able to assign them the correct labels, whereas other methods fail to accurately classifying all these images with the identical initially labeled images. Therefore we conclude that multi-modal curriculum learning is beneficial to ease the learning on complicated visual concepts.

C. Parametric Sensitivity

The weighting factor β and initial learning rate γ are two tuning parameters in our MMCL algorithm. This section stud-

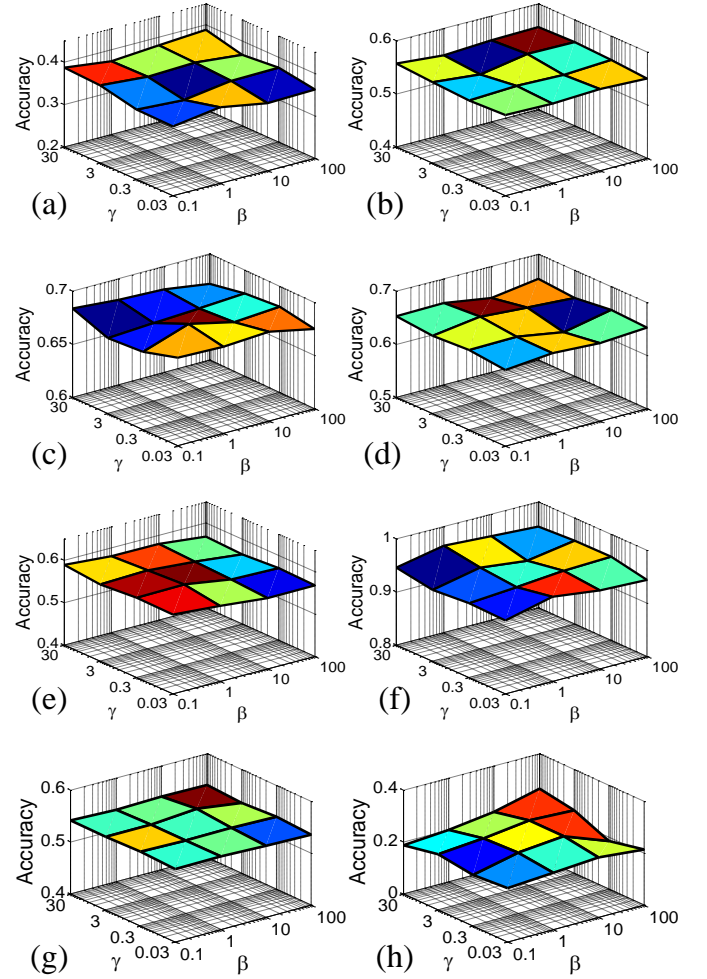


Fig. 7. Parametric sensitivity of MMCL w.r.t. β and γ . (a) is *CaltechAnimal* dataset, (b) is *Architecture* dataset, (c) is *MSRC* dataset, (d) is *UIUC* dataset, (e) is *Scene15* dataset, (f) is *ORLFace* dataset, (g) is *CIFAR100* dataset, and (h) is *NUS-WIDE* dataset.

ies how their variations influence the output accuracy. To this end, we fix the numbers of labeled images in the above eight datasets *CaltechAnimal*, *Architecture*, *MSRC*, *UIUC*, *Scene15*, *ORLFace*, *CIFAR100*, *NUS-WIDE* to 135, 375, 240, 400, 450, 120, 30000, 6720, respectively, and change β and γ to see the produced classification accuracy. The experimental results are illustrated in Fig. 7. It can be observed that even though β and γ cover wide ranges ($\beta \in [0.1, 100]$ and $\gamma \in [0.03, 30]$), the accuracies remain substantially unchanged, suggesting that the model output is very robust to the variations of these tuning parameters. As a result, the parameters incorporated by MMCL can be easily adjusted. Besides, MMCL is shown to achieve satisfactory performance overall on all datasets when $\beta = 10$ and $\gamma = 3$, which explains the reason that we choose this parameter setting for our experiments.

V. CONCLUSION

This paper proposed a novel curriculum learning approach, dubbed multi-modal curriculum learning, to optimize the quality of semi-supervised image classification. Benefiting from the wisdom of multiple teachers, the information from

different feature modalities is properly exploited and integrated, based on which a curriculum learning sequence (*i.e.*, a sequence for classifying unlabeled images) is generated in a simple-to-difficult order. Through extensive experiments, we demonstrated the superiority of the proposed multi-modal curriculum learning over the state-of-the-arts in terms of semi-supervised image classification accuracy. We also found that our approach is general in nature and hence readily applicable to other semi-supervised classification problems. In the future, we plan to extend MMCL to dealing with the noisy label cases [56], in which the labeled images with potentially incorrect labels are difficult and should be re-classified in later propagations.

REFERENCES

- [1] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *Computer Vision (ICCV), IEEE International Conference on*, 2013, pp. 1737–1744.
- [2] D. Dai and L. Gool, "Ensemble projection for semi-supervised image classification," in *Computer Vision (ICCV), IEEE International Conference on*, 2013, pp. 2072–2079.
- [3] R. Fergus, Y. Weiss, and A. Torralba, "Semi-supervised learning in gigantic image collections," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 522–530.
- [4] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2010, pp. 902–909.
- [5] Y. Luo, D. Tao, B. Geng, C. Xu, and S. Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classification," *Image Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 523–536, 2013.
- [6] L. Jing and M. Ng, "Sparse label-indicator optimization methods for image classification," *Image Processing, IEEE Transactions on*, vol. 23, no. 3, pp. 1002–1014, 2014.
- [7] X. Liu, T. Guo, L. He, and X. Yang, "A low-rank approximation-based transductive support tensor machine for semisupervised classification," *Image Processing, IEEE Transactions on*, vol. 24, no. 6, pp. 1825–1838, 2015.
- [8] B. Wang, Z. Tu, and J. Tsotsos, "Dynamic label propagation for semi-supervised multi-class multi-label classification," in *Computer Vision (ICCV), IEEE International Conference on*, 2013, pp. 425–432.
- [9] W. Xie, Z. Lu, Y. Peng, and J. Xiao, "Graph-based multimodal semi-supervised image classification," *Neurocomputing*, vol. 138, pp. 167–179, 2014.
- [10] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. International Conference on Machine Learning (ICML)*, 2009, pp. 41–48.
- [11] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," CMU-CALD-02-107, Tech. Rep., 2002.
- [12] X. Zhu and B. Goldberg, *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers, 2009.
- [13] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2005.
- [14] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. International Conference on Machine Learning (ICML)*, 1999, pp. 200–209.
- [15] Y. Li and Z. Zhou, "Towards making unlabeled data never hurt," in *Proc. International Conference on Machine Learning (ICML)*, 2011, pp. 1081–1088.
- [16] Y. Wang and S. Chen, "Safety-aware semi-supervised classification," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 11, pp. 1763–1772, 2013.
- [17] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.
- [18] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. International Conference on Machine Learning (ICML)*, 2003, pp. 912–919.
- [19] C. Gong, D. Tao, K. Fu, and J. Yang, "Fick's law assisted propagation for semisupervised learning," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 26, no. 9, pp. 2148–2162, 2015.
- [20] A. Shrivastava, S. Singh, and A. Gupta, "Constrained semi-supervised learning using attributes and comparative attributes," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 369–383.
- [21] A. Mahmood, A. Mian, and R. Owens, "Semi-supervised spectral clustering for image set classification," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2014, pp. 121–128.
- [22] G. Valls, T. Bandos, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 10, pp. 3044–3054, 2007.
- [23] C. Xu, T. Liu, D. Tao, and C. Xu, "Local rademacher complexity for multi-label learning," *Image Processing, IEEE Transactions on*, vol. 25, no. 3, pp. 1495–1507, 2016.
- [24] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv:1304.5634*, 2013.
- [25] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. Conference on Computational learning theory*, 1998, pp. 92–100.
- [26] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *Multimedia, IEEE Transactions on*, vol. 17, no. 11, pp. 2049–2058, 2015.
- [27] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *The Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.
- [28] M. White, X. Zhang, D. Schuurmans, and Y. Yu, "Convex multi-view subspace learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1673–1681.
- [29] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multiview matrix completion for multilabel image classification," *Image Processing, IEEE Transactions on*, vol. 24, no. 8, pp. 2355–2368, 2015.
- [30] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 1189–1197.
- [31] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2078–2086.
- [32] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann, "Self-paced curriculum learning," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.
- [33] C. Gong, D. Tao, W. Liu, L. Liu, and J. Yang, "Label propagation via teaching-to-learn and learning-to-teach," *Neural Networks and Learning Systems, IEEE Transactions on*, 2016.
- [34] C. Gong, D. Tao, J. Yang, and W. Liu, "Teaching-to-learn and learning-to-teach for multi-label propagation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.
- [35] Y. Lee and K. Grauman, "Learning the easy things first: Self-paced visual category discovery," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2011, pp. 1721–1728.
- [36] J. Supancic and D. Ramanan, "Self-paced learning for long-term tracking," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2013, pp. 2379–2386.
- [37] L. Jiang, D. Meng, T. Mitamura, and A. Hauptmann, "Easy samples first: self-paced reranking for zero-example multimedia search," in *ACM Multimedia Conference (ACM MM)*, 2014, pp. 547–556.
- [38] M. Karasuyama and H. Mamitsuka, "Manifold-based similarity adaptation for label propagation," in *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 1547–1555.
- [39] X. Zhu, J. Lafferty, and Z. Ghahramani, "Semi-supervised learning: From Gaussian fields to Gaussian processes," CMU-CS-03-175, Tech. Rep., 2003.
- [40] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
- [41] H. Qiu and E. Hancock, "Clustering and embedding using commute times," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 11, pp. 1873–1890, 2007.
- [42] D. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.
- [43] P. Absil and J. Malick, "Projection-like retractions on matrix manifolds," *SIAM Journal on Optimization*, vol. 22, no. 1, pp. 135–158, 2012.
- [44] F. Pompili, N. Gillis, P. Absil, and F. Glineur, "Two algorithms for orthogonal nonnegative matrix factorization with application to clustering," *arXiv:1201.0901v2*, 2014.
- [45] D. Zhou and O. Bousquet, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems (NIPS)*, 2003, pp. 321–328.

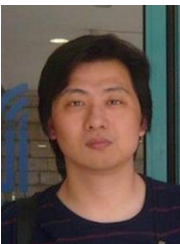
- [46] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 401–408.
- [47] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [48] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *European Conference on Computer Vision (ECCV)*. Springer, 2004, pp. 469–481.
- [49] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.
- [50] Z. Xu, Z. Hong, Y. Zhang, J. Wu, A. Tsoi, and D. Tao, "Multinomial latent logistic regression for image understanding," *Image Processing, IEEE Transactions on*, vol. 25, no. 2, pp. 973–987, 2016.
- [51] L. Li and F. Li, "What, where and who? classifying events by scene and object recognition," in *Computer Vision (ICCV), IEEE International Conference on*, 2007, pp. 1–8.
- [52] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2006, pp. 2169–2178.
- [53] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," U. Toronto, Tech. Rep., 2009.
- [54] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*. ACM, 2009, pp. 48–56.
- [55] M. Karasuyama and H. Mamitsuka, "Multiple graph label propagation by sparse integration," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 12, pp. 1999–2012, 2013.
- [56] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 3, pp. 447–461, 2016.



Dacheng Tao (F'15) is Professor of Computer Science with the Centre for Quantum Computation & Intelligent Systems, and the Faculty of Engineering and Information Technology in the University of Technology Sydney. He mainly applies statistics and mathematics to data analytics problems and his research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 200+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM'07, the best student paper award in IEEE ICDM'13, and the 2014 ICDM 10-year highest-impact paper award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award and the UTS Vice-Chancellor's Medal for Exceptional Research. He is a Fellow of the IEEE, OSA, IAPR and SPIE.



Stephen J. Maybank (F'11) received the BA degree in mathematics from King's College Cambridge in 1976 and the PhD degree in computer science from Birkbeck College, University of London in 1988. He was a research scientist at GEC from 1980 to 1995, first at MCCS, Frimley, and then, from 1989, at the GEC Marconi Hirst Research Centre in London. In 1995 he became a lecturer in the Department of Computer Science at the University of Reading and in 2004 he became a professor in the Department of Computer Science and Information Systems at Birkbeck College, University of London. His research interests include camera calibration, visual surveillance, tracking, filtering, applications of projective geometry to computer vision and applications of probability, statistics and information theory to computer vision. Maybank is the author or co-author of more than 160 scientific publications and one book. He is a Fellow of the IEEE and a Fellow of the Royal Statistical Society. He received the Koenderink Prize in 2008.



Wei Liu (M'14) received the Ph.D. degree from Columbia University, New York, NY, USA, in 2012. He was a recipient of the 2013 Jury Award for Best Thesis of Columbia University. He has been a research staff member of IBM T. J. Watson Research Center, Yorktown Heights, NY, USA since 2012. His research interests include machine learning, big data analytics, computer vision, pattern recognition, and information retrieval.



Chen Gong received his bachelor degree from East China University of Science and Technology (E-CUST) in 2010. Currently he is a dual doctoral degree student at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University (SJTU), and the Centre for Quantum Computation & Intelligent Systems, University of Technology Sydney (UTS), under the supervision of Prof. Jie Yang and Prof. Dacheng Tao. His research interests mainly include machine learning, data mining and learning-based vision problems. He has published 28

technical papers at prominent journals and conferences such as IEEE T-NNLS, IEEE T-IP, IEEE T-CYB, CVPR, AAAI, ICME, etc. He received the "National Scholarship" awarded by the Ministry of Education in 2013 and 2014, the "Excellent Self-financed Overseas Student Scholarship" awarded by China Scholarship Council in 2015, and the "IBM Excellent Student Scholarship" in 2015.



Guoliang Kang received his bachelor degree of Automation at Chongqing University in 2011, and his Master degree of Pattern Recognition and Intelligent Systems at Beihang University in 2014. He is currently a PhD candidate in the University of Technology Sydney. His research interests include machine learning and computer vision.



Jie Yang received his PhD from the Department of Computer Science, Hamburg University, Germany, in 1994. Currently, he is a professor at the Institute of Image Processing and Pattern recognition, Shanghai Jiao Tong University, China. He has led many research projects (*e.g.*, National Science Foundation, 863 National High Tech. Plan), had one book published in Germany, and authored more than 200 journal papers. His major research interests are object detection and recognition, data fusion and data mining, and medical image processing.